206–211 Nucleic Acids Research, 1997, Vol. 25, No. 1

IMGT, the international ImMunoGeneTics database

Véroniqu Giudicelli, D nys Chaume[1], Julia Bodmer[2], Werner Müll r[3], Chantal Busin, Steven Marsh[2], Ronald Bontrop[4], Lemaitre Marc[5], Ansar Malik[6] and Marie-Paule Lefranc*

Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UMR CNRS 5535, BP5051, 1919 route de Mende, 34033 Montpellier Cedex 1, France, [1]CNUSC, Montpellier, France, [2]ICRF, London, UK, [3]IFG, Köln, Germany, [4]BPRC, Rijswijk, The Netherlands, [5]EUROGENTEC S.A., Seraing, Belgium and [6]EMBL Outstation EBI, Hinxton, UK

Received August 21, 1996; Accepted September 19, 1996

ABSTRACT

IMGT, the international ImMunoGeneTics database, is an integrated database specializing in immunoglobulins, T-cell receptors (TcR) and major histocompatibility complex (MHC) of all vertebrate species, initiated and co-ordinated by Marie-Paule Lefranc, CNRS, Montpellier II University, Montpellier, France (lefranc@ligm.crbm.cnrs-mop.fr). IMGT includes two databases: LIGM-DB (for immunoglobulins and TcR) and MHC/HLA-DB. IMGT comprises expertly annotated sequences and alignment tables. LIGM-DB contains more than 19 000 immunoglobulin and TcR sequences from 78 species. MHC/HLA-DB contains class I and class II human leukocyte antigen alignment tables. An IMGT tool, DNAPLOT, developed for immunoglobulins, TcR and MHC sequence alignments, is also available. IMGT works in close collaboration with the EMBL database. IMGT goals are to establish a common data access to all immunogenetics data, including sequences, oligonucleotide primers, gene maps and other genetic data of immunoglobulins, TcR and MHC molecules, and to provide a graphical user-friendly data access. IMGT will have important implications in medical research (repertoire in autoimmune diseases, AIDS, leukemias, lymphomas), therapeutical approaches (antibody engineering), genome diversity and genome evolution studies. IMGT can be accessed at http://imgt.cnusc.fr:8104 and http://www.ebi.ac.uk/IMGT

INTRODUCTION

The molecular synthesis of the immunoglobulin and T-cell receptor (TcR) chains (1,2) is particularly complex and unique as it includes biological mechanisms such as DNA molecular rearrangements in seven loci (three for immunoglobulins and four for TcR) located on four different chromosomes in human, nucleotide deletions and insertions at the rearrangement junctions, and hypermutations in the immunoglobulin loci. The number of potential protein forms of immunoglobulins and TcR is almost unlimited. Owing to the complexity and high number of published sequences, data control and detailed annotations are a very difficult task for the generalist databanks: EMBL (3), GenBank (4), DDBJ. Furthermore, until now, only poor efforts have been made to standardize the description of the immunoglobulin and TcR sequences at the nucleotide or protein level. Only few feature labels are specifically used in generalist databases for immunoglobulin and TcR annotations (seven in EMBL) and this often leads to errors or misinterpretations. These observations together with the proposal made by Fuchs and Cameron (5) to create specialized databases in collaboration with the generalist databases, were the starting point of IMGT in 1992 (6) (see Fig. 1 for the IMGT home page at http://imgt.cnusc.fr:8104). Before the physical implementation of the database, the main and the longest objective was to establish rules for describing immunoglobulin and TcR sequences of any species. This was the major foundation for a consistent expertise.

IMGT RULES

Standardization of keywords

IMGT keywords for immunoglobulins and TcR comprise the following. (i) *General keywords*. Indispensable for the sequence assignments, they are described in an exhaustive and non-redundant list, and are organized in a tree structure. (ii) *Specific keywords*. They are more specifically associated with particularities of the sequences (orphon, pseudogene...) or to diseases (leukemia, lymphoma, tumor...). The list is not definitive and new specific keywords can easily be added if needed.

The whole list of keywords can be reached using WWW browser at the URL http://imgt.cnusc.fr:8104/textes/LECT/kw.html.

Standardization of sequence annotation

Immunoglobulin and TcR sequences have been analyzed at the DNA and protein level in order to define a list of labels for the structural and functional motifs. More than 160 labels were shown to be necessary for an accurate annotation. The annotation is the most critical step and a very time-consuming process as about 50 sequences a week can be annotated by an experienced annotator. Levels of annotation have been defined, which allow the users to query sequences in IMGT/LIGM-DB even though they are not fully

*To whom correspondence should be addressed. Tel: +33 467 61 36 34; Fax: +33 467 04 02 31; Email: lefranc@ligm.crbm.cnrs-mop.fr

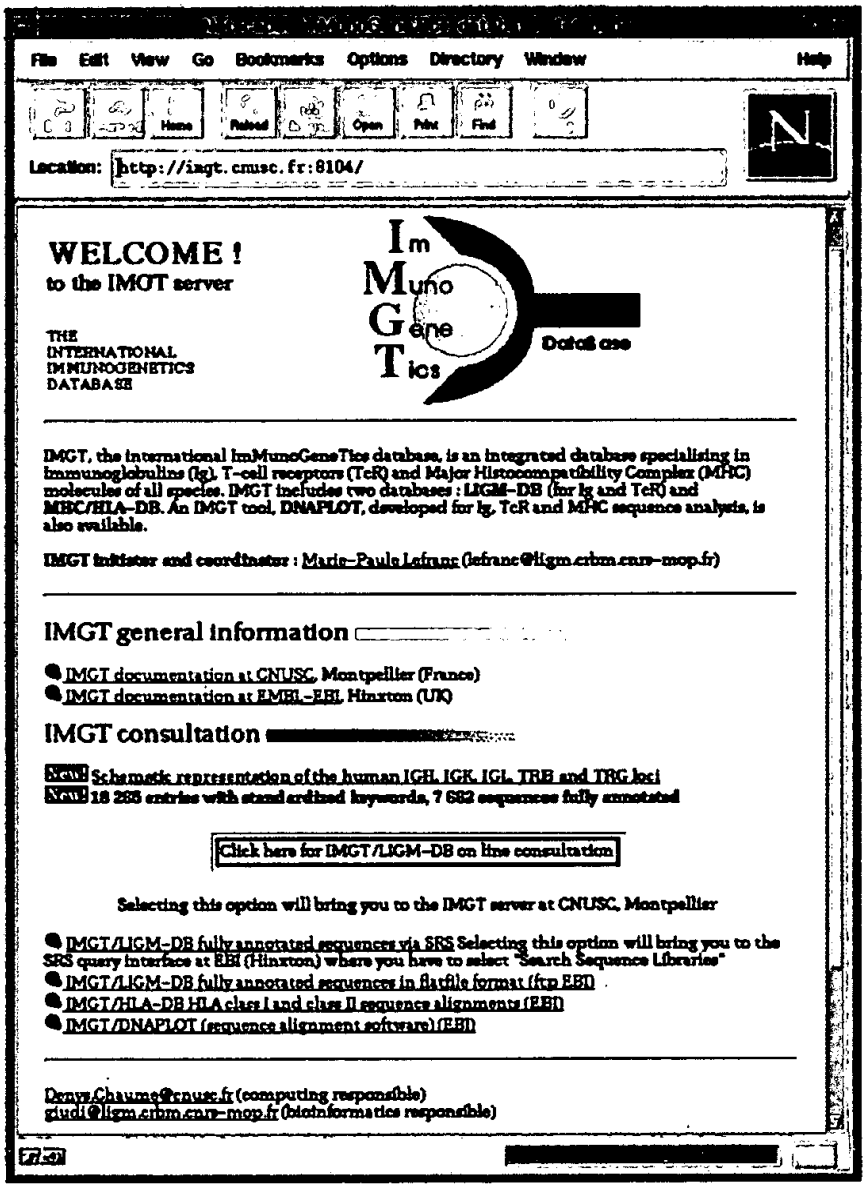

Figure 1. The IMGT international ImMunoGeneTics database WWW home page.

annotated. The list of labels with their corresponding definition and main schemas are available at the URL: http://imgt.cnusc.fr:8104/textes/LECT/labeldef.html (Fig. 2).

Standardization of immunoglobulin and TcR gene designation

The objective is to provide immunologists and geneticists with a unique nomenclature per locus which will allow extraction and comparison of data for the complex B- and T-cell antigen receptor molecules, whatever the species. In a first step, data concerning the human immunoglobulin and TcR genes have been standardized and maps of loci with IMGT nomenclature, correspondence to other gene designations and gene functionality are available from the IMGT home page at http://imgt.cnusc.fr:8104, since August 1996 (Fig. 3). These maps will be completed by tables.

IMGT/LIGM-DB ORGANIZATION AND CONTENT

LIGM-DB development is mainly based on a relational model organization. The database is maintained with SYBASE as relational DBSM (Data Base System Manager) on Unix IBM workstation at CNUSC (Centre National Universitaire Sud de Calcul) in Montpellier (France). CNUSC is in charge of the

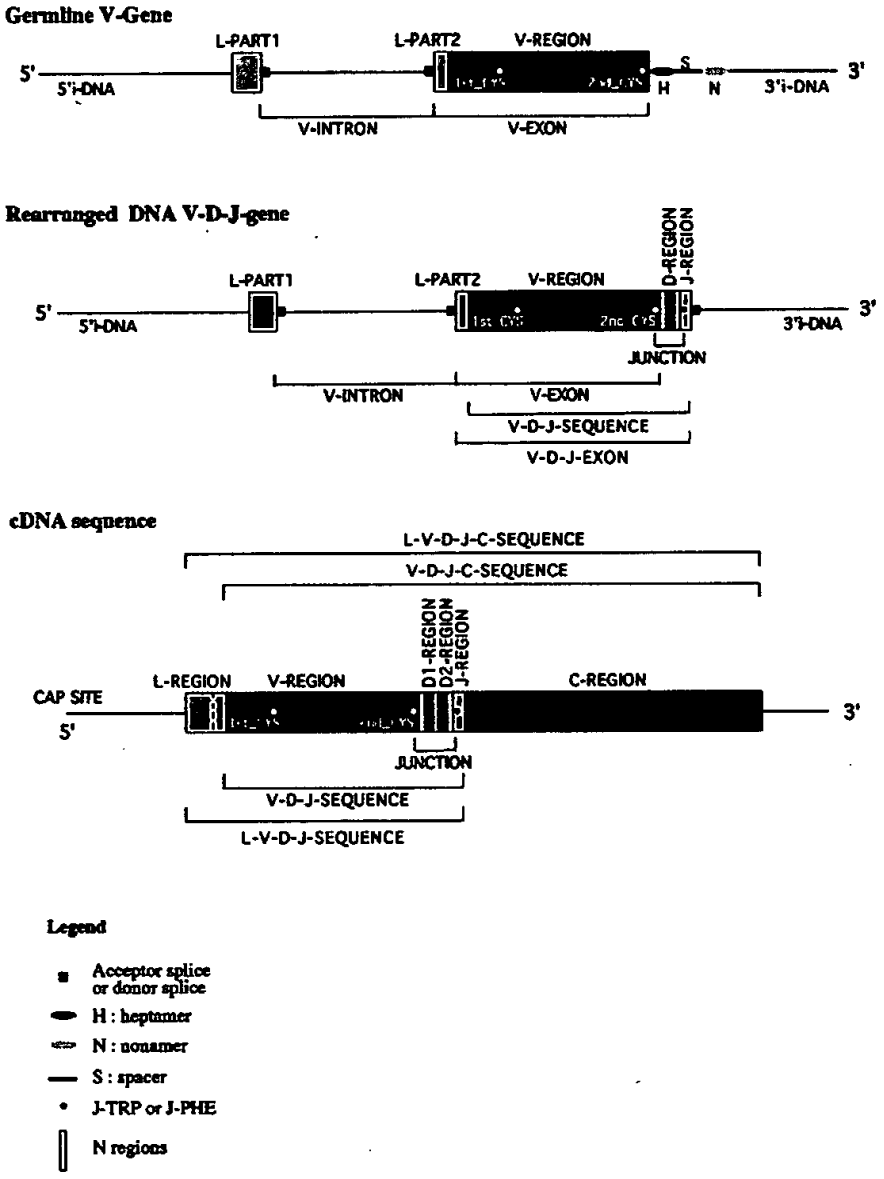

Figure 2. An example of graphical representation of labels defined in IMGT/LIGM-DB.

computing exploitation. New releases of the relational schema and updates of the database structure, that closely follow the results of biological research, are under LIGM and CNUSC responsibility.

In November 1996, LIGM-DB contained 19 540 nucleic acid sequences of immunoglobulins or TcR from 78 species. IMGT sequences are identified by the EMBL accession number. IMGT data comprise core data that consist of sequence data, bibliographical references, taxonomic data retrieved from EMBL entries, completed with annotations, specific analysis and expertise provided by LIGM. IMGT/LIGM-DB standardized keywords have been assigned to all entries, and 7908 sequences are now fully annotated. Since August 1996, the IMGT/LIGM-DB content follows closely the immunoglobulin and TcR EMBL one, with the advantage of being deleted from sequences which have previously been wrongly assigned to immunoglobulins and TcR.

DATA COLLECTION AND ANNOTATION

Source of data

The unique source of data is the generalist database EMBL. Once the sequences are allowed by the authors to be made public, EMBL sends automatically immunoglobulin and TcR sequences

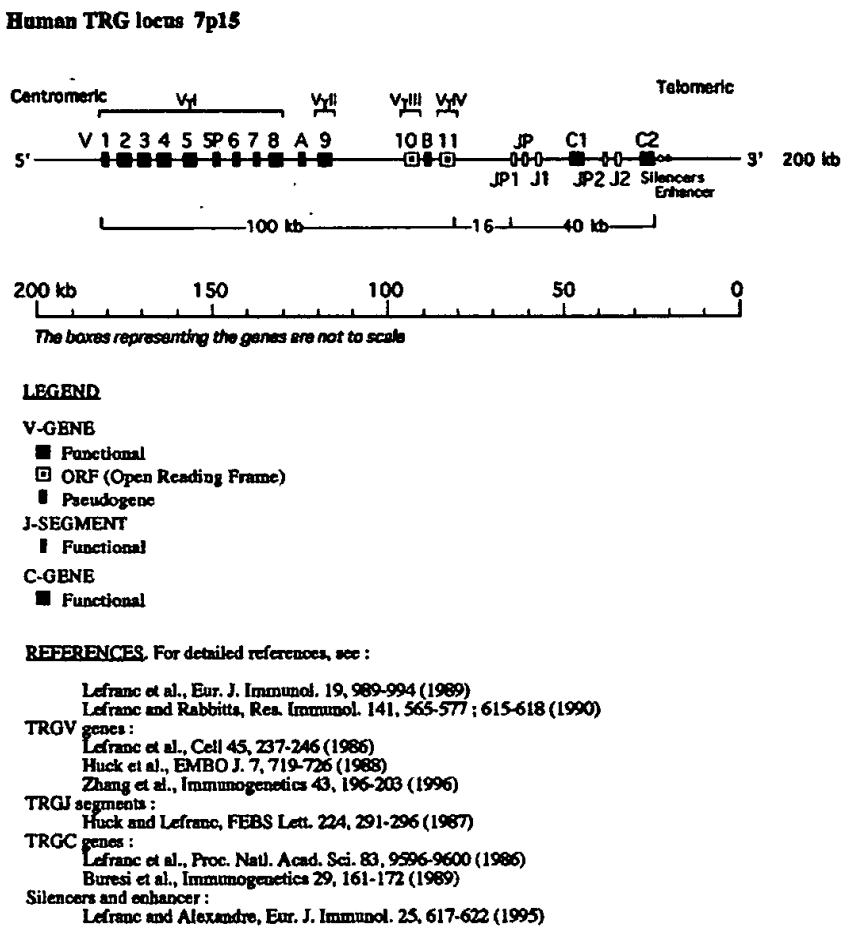

Figure 3. An example of map representation of a TcR locus.

to LIGM by mail. After control by LIGM curators, sequences are scanned in order to store IMGT non-specific information, such as bibliographical references and taxonomic data.

Keyword assignment

Standardized keywords are so far assigned manually to each new sequence by LIGM annotators. Procedure for the automation of the IMGT keyword attribution is in development.

Annotation procedure

The annotation of sequences is the most limiting step in the expertise of the data. Several approaches have been developed in order to increase the number of annotated sequences per month, and efforts are currently done to improve LIGM efficiency in this field.

Automatic motif recognition. The C written general algorithm for motif searches, BioMotif, developed by the Laboratoire de Physique Mathématique of Montpellier, France, has been specifically adapted for immunoglobulin and TcR sequences. This algorithm, designated as LIGMotif and based on the use of EMBL flat files, scans the nucleic acid sequence for immunoglobulin or TcR specific motifs (characteristic amino acids in conserved positions...), according to the presence of information such as receptor and chain type. At the end of the search, it provides a text file which contains potential solutions for delimitation of functional or structural subregions. It also provides the FR (FRamework) and CDR (Complementarity Determining Region) delimitations (1).

Annotations in delayed conditions. In order to make the annotators independent from Internet connection and allow them to annotate 'in any place', we have developed a simple text mode release of the annotation module that facilitates the data acquisition on any local computer. Resulting annotations are then sent by mail, ftp or tape to LIGM. Annotators can also use text files resulting from LIGMotif analysis and, after control of the annotations, include them into IMGT.

A tool for immunoglobulin, TcR and MHC sequence alignments: DNAPLOT. Immunologists mainly use sequence comparison either to search similar or identical sequences in databases, or to classify immunoglobulins and TcR sequences in subgroups, in which the sequences share more than 75% similarity and can be detected by the same probe. The Institut Für Genetic (IFG) of

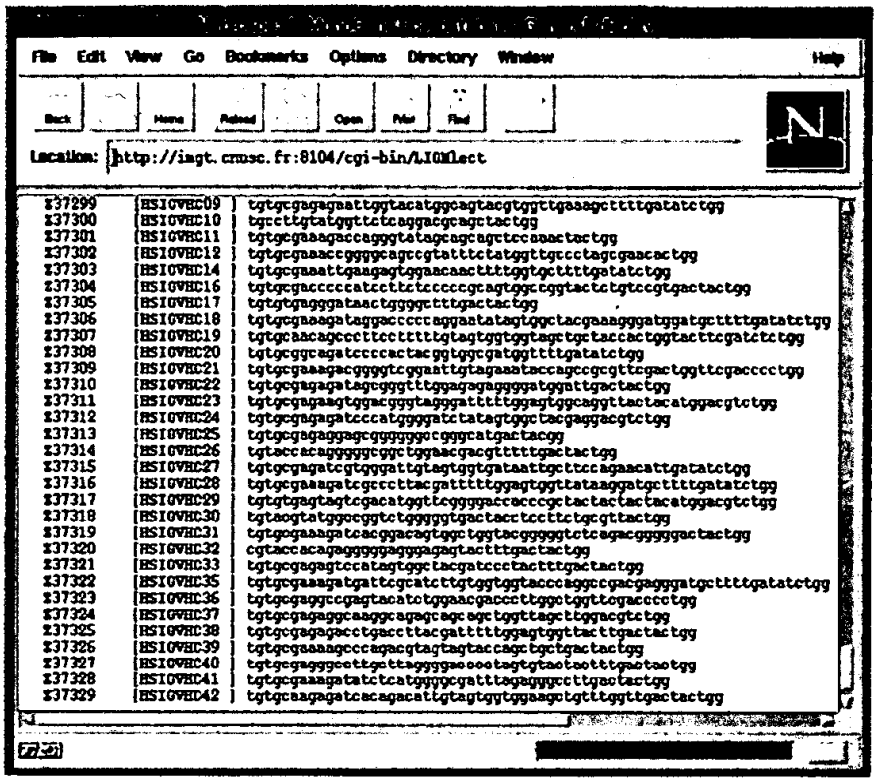

Figure 4. List of specific coding regions ('JUNCTION') extracted from an example of query resulting sequences.

Köln, Germany, has developed a program DNAPLOT which generates, displays and analyzes nucleotide sequence alignments. DNAPLOT is complementary to existing programs, such as GDE, CLUSTALW, FASTA, BLAST or READSEQ, and does not replace their functions. It can also propose assignment of rearranged or expressed variable genes to the potential germline genes. DNAPLOT is available at:
http://www.genetik.uni-koeln.de/dnaplot/
and from the IMGT Home page.

DATA DISTRIBUTION

No restrictions are placed on the use or redistribution of the IMGT data. Currently, IMGT is available through Internet and on the quaterly CD-ROM distributed by the EMBL data library.

Flat file production

Flat files are produced in collaboration with EBI. Names of entries remain the EMBL accession number. IMGT/LIGM-DB flat file typical entries provide LIGM expertise: standardized LIGM keywords appear in KW code lines, complement to definition in DE lines and sequence description with LIGM labels in FT code lines. Core data, as well as cross-references to other databases in DR lines are kept from EMBL. Flat file format allows IMGT/LIGM-DB data to be compatible with the most efficient software for information retrieval, data manipulation such as the largely distributed browser SRS, which also allows consultation of the cross-referenced databases (available at http://www.ebi.ac.uk/srs/srsc). IMGT/LIGM-DB flat files are available on EMBL anonymous ftp server (ftp.ebi.ac.uk in pub/databases/imgt) and are also distributed with many other databases on the EMBL CD-ROM.

Interactive access to IMGT on the WWW

A WWW IMGT server has been installed at CNUSC and can be reached with Mosaic and Netscape WWW browser at the URL http://imgt.cnusc.fr:8104. The biologist needs were taken into account for the development of the interface WWW–SYBASE which allows users to create very specific and structured queries combining aspects of relational database and hypertext. Requests can be performed through distinct modules that allow to classify search criteria type. At the issue of a run, a number of resulting sequences is proposed and it is then possible to either look at the solutions, or to add new conditions to modify the results, keeping in memory the previously selected criteria. There are several ways to retrieve the results, in particular it is possible to extract specific coding regions from the query resulting sequences even though alignment tools are not yet integrated into IMGT (Fig. 4). Links with Medline are now available.

CONCLUSIONS

IMGT is developed by LIGM (Montpellier, France) in collaboration with CNUSC (Montpellier, France), EMBL-EBI (Hinxton,

UK), ICRF (London, UK), IFG (Köln, Germany), BPRC (Rijswijk, The Netherlands) and EUROGENTEC S.A. (Seraing, Belgium). The information provided by IMGT is of much value to clinicians and biological scientists in general. The main objectives for the next 3 years include the development of a WWW interface for direct submission of the data by the authors, development of MHC/HLA-DB and extension to all species. New specific databases will be developed and integrated into IMGT: a protein database for immunoglobulins and TcR which will contain translations of potentially functional and ORF sequences from LIGM-DB, and protein data from Kabat (7) and SWISS_PROT (8), and an oligonucleotide primer database for immunoglobulins, TcR and MHC. IMGT will include, in the future, analysis of genetics data and displays of physical maps. IMGT is designed to allow common access to all immunogenetics data. This approach is based on a very tight collaboration with EMBL for the nucleotide sequence data, with SWISS-PROT for the protein sequence data and with IGD for providing a user friendly interface for the mapping and genetic data. Particular attention will be given to the establisment of cross-referencing links to other databases pertinent to the users of IMGT.

ACCESS AND CONTACT

CNUSC WWW server at http://imgt.cnusc.fr:8104. Contact Denys.Chaume@cnusc.fr.

EBI servers at http://www.ebi.ac.uk/imgt; ftp.ebi.ac.uk (folder/pub/databases/imgt); contact malik@ebi.ac.uk

For comments and suggestions contact giudi@ligm.crbm.cnrs-mop.fr

IMGT initiator and coordinator: Marie-Paule Lefranc, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UMR CNRS 5535, BP5051, 1919 route de Mende, 34033 Montpellier Cedex 1, France; Tel: +33 467 61 36 34; Fax: +33 467 04 02 31; Email: lefranc@ligm.crbm.cnrs-mop.fr

ACKNOWLEDGMENTS

We thank Gérard Mennessier for the development of LIGMotif. We are deeply grateful to Valérie Barbié, Anne Bouisson, Géraldine Folch, Sophie Lefebvre, Nathalie Pallares and Gaëlle Rousseaux who are the present LIGM-DB annotators. IMGT is funded by the European Union's BIOMED1 and BIOTECH programmes, the CNRS (Centre National de la Recherche Scientifique), and the MENESR (Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche). Subventions have been received from Association pour la Recherche sur le Cancer, Association de Recherche sur la Polyarthrite, Fondation pour la Recherche Médicale, Groupement de Recherche et d'Etude sur les Génomes and the Région Languedoc-Roussillon.

REFERENCES

1 Honjo, T. and Alt, F.W. (1995) *Immunoglobulin genes.* Academic Press pp. 3–443.
2 Lefranc, M-P. (1990) *Eur. Cytokine Network*, **1**, 121–130.
3 Rodriguez-Tome, P., Stochr, P.J., Cameron, G.N. and Flores, T.P. (1996) *Nucleic Acids Res.*, **24**, 6–12.
4 Benson, D.A., Bouski, M., Lipman, D.J. and Ostell, J. (1996) *Nucleic Acids Res.*, **24**, 1–5.
5 Fuchs, R. and Cameron, G.N. (1991) *Prog. Biophysics Mol. Biol.*, **56**, 215–245.
6 Lefranc, M-P., Giudicelli, V., Busin, C., Malik, A., Mougenot, I., Déhais, P. and Chaume, D. (1995) *Ann. N. Y. Acad. Sci.*, **764**, 47–49.
7 Kabat, E.A., Wu, T.T., Perry, H.M., Gottesman, K.S. and Foeller, C. (1991) *Sequences of proteins of immunological interest.* National Institutes of Health, Bethesda.
8 Bairoch, A. and Apweiler, R. (1996) *Nucleic Acids Res.*, **24**, 21–25.